

Mathmatial Foundations of Reinforcement Learning

Model-free RL: Q-learning



Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

Fall 2023

Outline

Synchronous Q-learning

Asynchronous Q-learning

Bellman's optimality principle

Bellman operator

$$\mathcal{T}(Q)(s, a) := \underbrace{\mathbb{E}[r(s, a)]}_{\text{immediate reward}} + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[\underbrace{\max_{a' \in \mathcal{A}} Q(s', a')}_{\text{next state's value}} \right]$$

- one-step look-ahead

Bellman's optimality equation: Q^* is the *unique* fixed point to

$$\mathcal{T}(Q^*) = Q^*.$$

γ -contraction:

$$\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

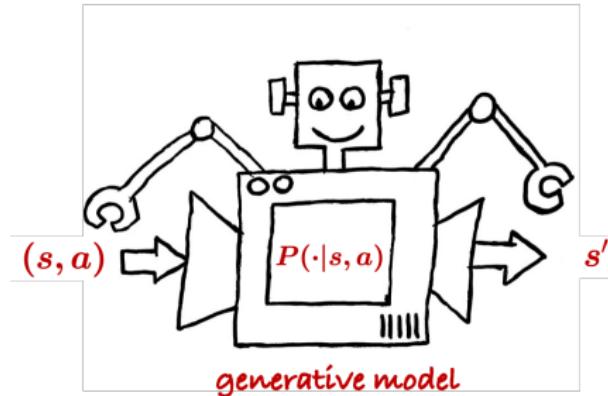


Richard Bellman

Synchronous Q-learning

Synchronous sampling with a generative model

— [Kearns and Singh, 1999]



For each state-action pair (s, a) , at each time t collect

$$(s, a, s')$$

Question: How many samples are necessary and sufficient to learn the optimal policy without worrying about exploration?

Q-learning: a classical model-free algorithm



Chris Watkins



Peter Dayan

Stochastic approximation for solving the **Bellman equation**

Robbins & Monro, 1951

$$Q = \mathcal{T}(Q)$$

where

$$\mathcal{T}(Q)(s, a) := \underbrace{r(s, a)}_{\text{immediate reward}} + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[\underbrace{\max_{a' \in \mathcal{A}} Q(s', a')}_{\text{next state's value}} \right].$$

Q-learning: a classical model-free algorithm



Chris Watkins



Peter Dayan

Q-learning [Watkins and Dayan, 1992] proceeds as

$$\begin{aligned} Q_{t+1}(s, a) &= \underbrace{(1 - \eta_t)Q_t(s, a) + \eta_t \mathcal{T}_t(Q_t)(s, a)}_{\text{draw the transition } (s, a, s') \text{ for all } (s, a)}, \quad t \geq 0 \\ &= Q_t(s, a) + \eta_t \left(r(s, a) + \gamma \max_{a'} Q_t(s', a') - Q_t(s, a) \right) \end{aligned}$$

$$\mathcal{T}_t(Q)(s, a) = r(s, a) + \gamma \max_{a'} Q(s', a')$$

$$\mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[\max_{a'} Q(s', a') \right]$$

Asymptotic convergence

Theorem 1 ([Watkins and Dayan, 1992])

Q-learning converges to the optimal Q-function Q^ asymptotically with probability 1 as long as*

$$\sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty.$$

- The first condition asks the learning rates to be not too small, while the second condition ensures that they are not too large.
- Many choices of learning rates satisfy this assumption.

What about the finite-time convergence rate of Q-learning?

Prior art: achievability

Question: How many samples are needed for $\|\hat{Q} - Q^*\|_\infty \leq \varepsilon$?

paper	learning rates	sample complexity
Even-Dar & Mansour '03	linear: $\frac{1}{t}$	$2^{\frac{1}{1-\gamma}} \frac{ \mathcal{S} \mathcal{A} }{(1-\gamma)^4 \varepsilon^2}$
Beck & Srikant '12	constant: $\frac{(1-\gamma)^4 \varepsilon^2}{ \mathcal{S} \mathcal{A} }$	$\frac{ \mathcal{S} ^2 \mathcal{A} ^2}{(1-\gamma)^5 \varepsilon^2}$
Wainwright '19	rescaled linear: $\frac{1}{1+(1-\gamma)t}$	$\frac{ \mathcal{S} \mathcal{A} }{(1-\gamma)^5 \varepsilon^2}$
Chen et al. '20	rescaled linear: $\frac{1}{\frac{1}{(1-\gamma)^2} + (1-\gamma)t}$	$\frac{ \mathcal{S} \mathcal{A} }{(1-\gamma)^5 \varepsilon^2}$
Chen et al. '20	constant: $(1-\gamma)^4 \varepsilon^2$	$\frac{ \mathcal{S} \mathcal{A} }{(1-\gamma)^5 \varepsilon^2}$

A note on the learning rates

Observation: the learning rate schedule $\eta_t = \frac{1}{t}$ leads to a sample complexity that scales **exponentially** in $\frac{1}{1-\gamma}$.

- Consider the following MDP with a single state $s = 1$, a single action $a = 1$, and $r(1, 1) = 1, P(1|1, 1) = 1$. Hence,

$$Q^*(1, 1) = \frac{1}{1 - \gamma}.$$

- The update rule of Q-learning with learning rate $\eta_t = \frac{1}{t}$ gives

$$\begin{aligned} Q_t(1, 1) &= \left(1 - \frac{1}{t}\right) Q_{t-1}(1, 1) + \frac{1}{t} (1 + \gamma Q_{t-1}(1, 1)) \\ &= \left(1 - \frac{1-\gamma}{t}\right) Q_{t-1}(1, 1) + \frac{1}{t}, \end{aligned}$$

A note on the learning rates - continued

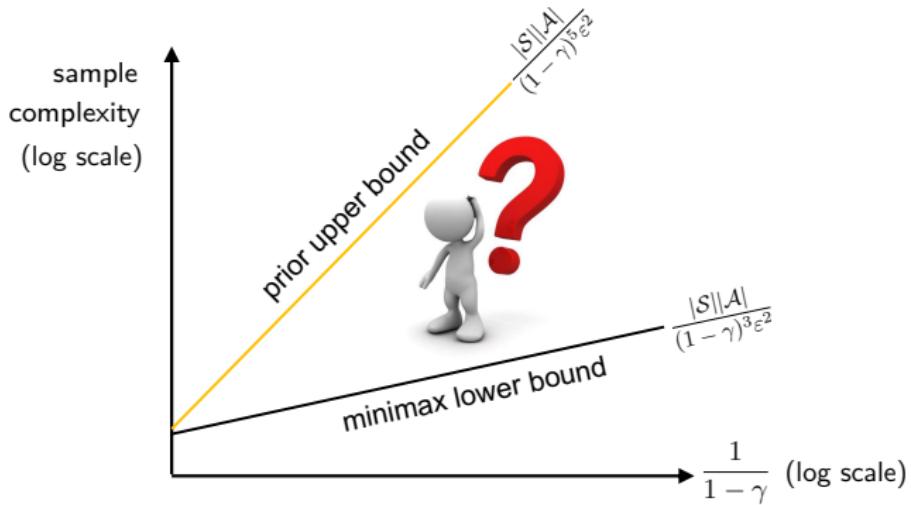
- From simple recursive relations, one can easily check that: when $\gamma \rightarrow 1$ and t is not too large, one has

$$\begin{aligned} Q_t - Q^* &= \prod_{i=1}^t \left[1 - \frac{1-\gamma}{i} \right] \cdot [Q_0 - Q^*] \\ &\approx \left[1 - \sum_{i=1}^t \frac{1-\gamma}{i} \right] \cdot [Q_0 - Q^*] \\ &\approx [1 - (1-\gamma) \log t] \cdot [Q_0 - Q^*]. \end{aligned}$$

This essentially implies that one needs to have $t \gtrsim 2^{O(\frac{1}{1-\gamma})}$ iterations to achieve $|Q_t - Q^*| < \frac{1}{2}|Q_0 - Q^*|$.

Consequently, the **rescaled** linear learning rates or constant learning rates provide better alternatives.

Can we close the gap?



All prior results require sample size of at least $\frac{|S||\mathcal{A}|}{(1-\gamma)^5 \varepsilon^2}$!

Is Q-learning sub-optimal, or is it an analysis artifact?

A sharpened sample complexity of Q-learning

Theorem 2 ([Li et al., 2021])

For any $0 < \varepsilon \leq 1$, Q-learning yields

$$\|\hat{Q} - Q^*\|_\infty \leq \varepsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right).$$

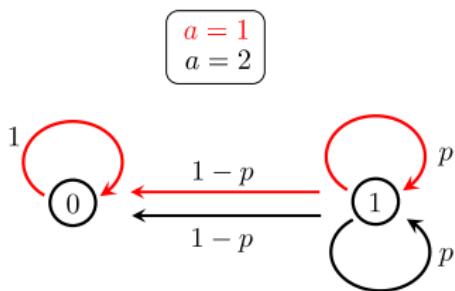
- Improves dependency on effective horizon $\frac{1}{1-\gamma}$
- Allows both constant and rescaled linear learning rate:

$$\frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \leq \eta_t \leq \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}}$$

A curious numerical example

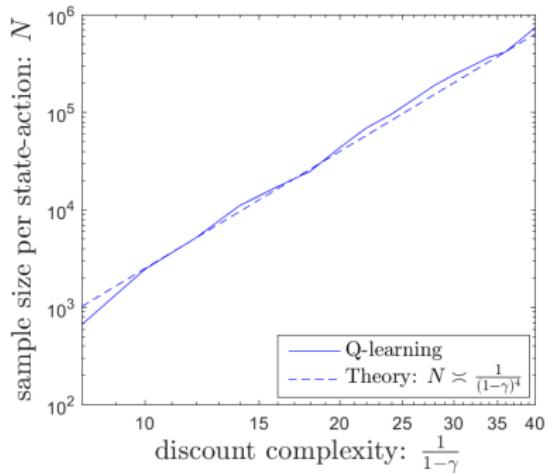
Numerical evidence: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4 \varepsilon^2}$ samples seem necessary . . .

— observed in [Wainwright, 2019a]



$$p = \frac{4\gamma - 1}{3\gamma}$$

$$r(0, 1) = 0, \quad r(1, 1) = r(1, 2) = 1$$



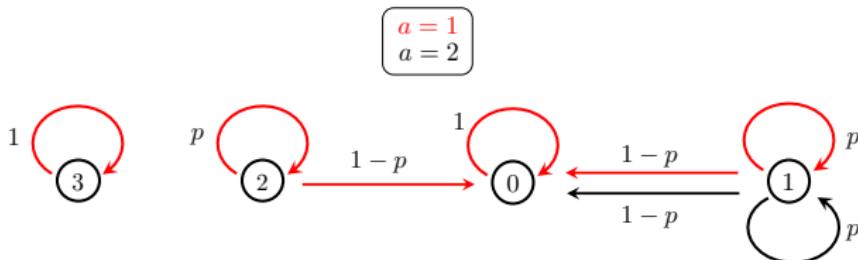
Q-learning is not minimax optimal

Theorem 3 ([Li et al., 2021])

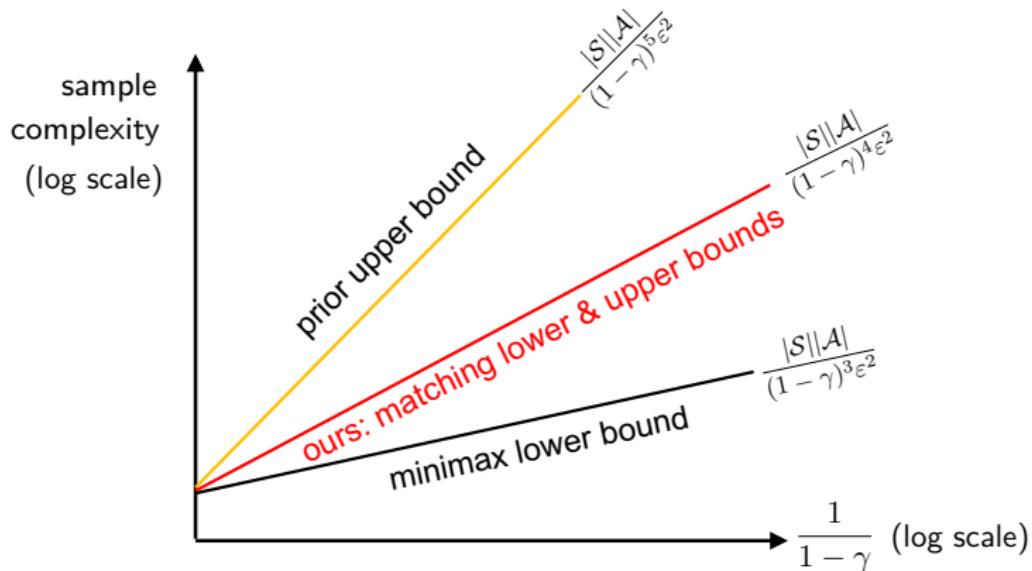
For any $0 < \varepsilon \leq 1$, there exist an MDP such that to achieve $\|\widehat{Q} - Q^*\|_\infty \leq \varepsilon$, Q-learning needs at least a sample complexity of

$$\widetilde{\Omega}\left(\frac{|S||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right).$$

- Tight **algorithm-dependent** lower bound
- Holds for both constant and rescaled linear learning rates



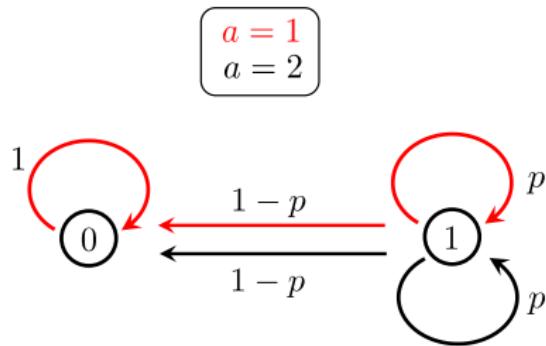
Where we stand now



Q-learning requires a sample size of $\frac{|S||\mathcal{A}|}{(1-\gamma)^4 \varepsilon^2}$.

Why is Q-learning sub-optimal?

Over-estimation of Q-functions [Thrun and Schwartz, 1993, Hasselt, 2010]:



- $\max_{a \in \mathcal{A}} \mathbb{E}X(a)$ tends to be over-estimated (high positive bias) when $\mathbb{E}X(a)$ is replaced by its empirical estimates using a small sample size.

Why is Q-learning sub-optimal?

The over-estimation of Q-functions often gets **worse** with a large number of actions [Van Hasselt et al., 2016].

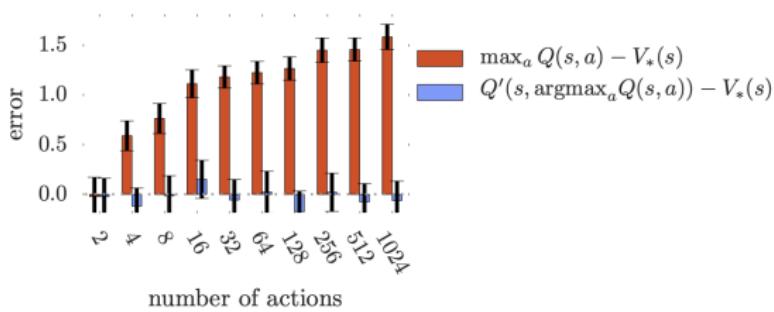


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s, a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q' , used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.

Double Q-learning

To mitigate the impact of over-estimation, [Hasselt, 2010] proposed **double Q-learning**, which uses two Q-estimates and updates one of them randomly at each round:

$$Q^1(s, a) = (1 - \eta_t)Q^1(s, a) + \eta_t \left(r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q^2(s', a') \right),$$

or

$$Q^2(s, a) = (1 - \eta_t)Q^2(s, a) + \eta_t \left(r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q^1(s', a') \right).$$

- Decouple the randomness in value updates and action selection.
- Empirically very successful when integrated with deep RL
[Van Hasselt et al., 2016].

TD-learning: when the action space is a singleton



Richard Sutton

Stochastic approximation for solving Bellman equation $V = \mathcal{T}(V)$

$$\begin{aligned} V_{t+1}(s) &= (1 - \eta_t)V_t(s) + \eta_t \mathcal{T}_t(V_t)(s) \\ &= V_t(s) + \eta_t \underbrace{\left[r(s) + \gamma V_t(s') - V_t(s) \right]}_{\text{temporal difference}}, \quad t \geq 0 \end{aligned}$$

$$\mathcal{T}_t(V)(s) = r(s) + \gamma V(s')$$

$$\mathcal{T}(V)(s) = r(s) + \gamma \mathbb{E}_{s' \sim P(\cdot | s)} V(s')$$

Sample complexity of TD-learning

Theorem 4 ([Li et al., 2021])

For any $0 < \varepsilon \leq 1$, TD-learning yields

$$\|\hat{V} - V^*\|_\infty \leq \varepsilon$$

with sample complexity at most

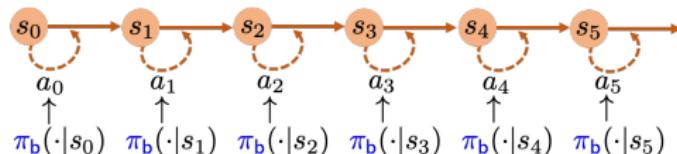
$$\tilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right).$$

- Near minimax-optimal (matches the minimax lower bound when the action space is a singleton) without the need of averaging or variance reduction.
- Allows both constant and rescaled linear learning rate.

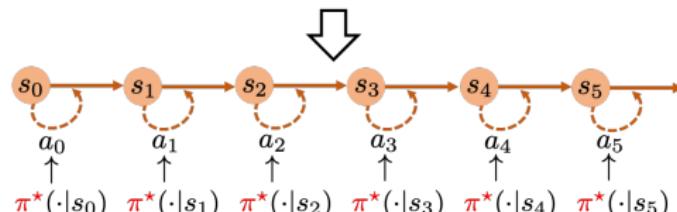
Asynchronous Q-learning

Markovian samples and behavior policy

observed:



learn:

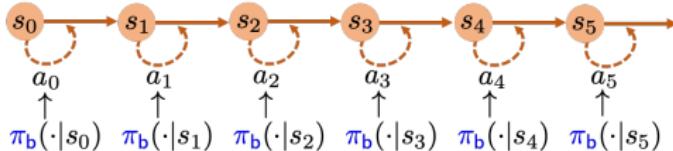


Observed: $\underbrace{\{s_t, a_t, r_t\}_{t \geq 0}}_{\text{stationary Markovian trajectory}}$ generated by **behavior policy** π_b

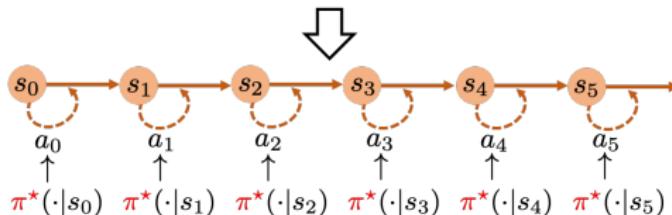
Goal: learn optimal value V^* and Q^* based on sample trajectory

Key quantities of sample trajectory

observed:



learn:



- minimum state-action occupancy probability (**uniform coverage**)

$$\mu_{\min} := \min \underbrace{\mu_{\pi_b}(s, a)}_{\text{stationary distribution}}$$

- mixing time: t_{mix} , which captures the time to reach the steady state

Asynchronous Q-learning



Chris Watkins



Peter Dayan

Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

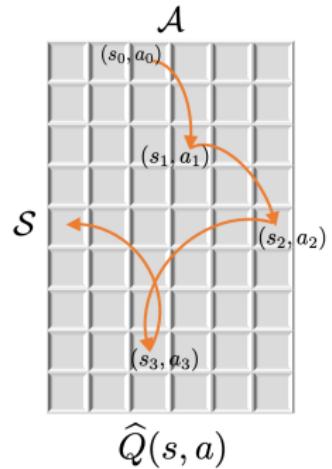
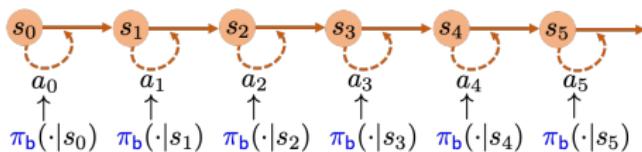
$$\underbrace{Q_{t+1}(s_t, a_t) = (1 - \eta_t)Q_t(s_t, a_t) + \eta_t \mathcal{T}_t(Q_t)(s_t, a_t)}_{\text{only update } (s_t, a_t)\text{-th entry}}, \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s_t, a_t) = r(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a')$$

$$\mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[\max_{a'} Q(s', a') \right]$$

Q-learning on Markovian samples

observed:



- **asynchronous:** only a single entry is updated each iteration
 - resembles Markov-chain *coordinate descent*
- **off-policy:** target policy $\pi^* \neq$ behavior policy π_b

Sample complexity of asynchronous Q-learning

Theorem 5 ([Li et al., 2022])

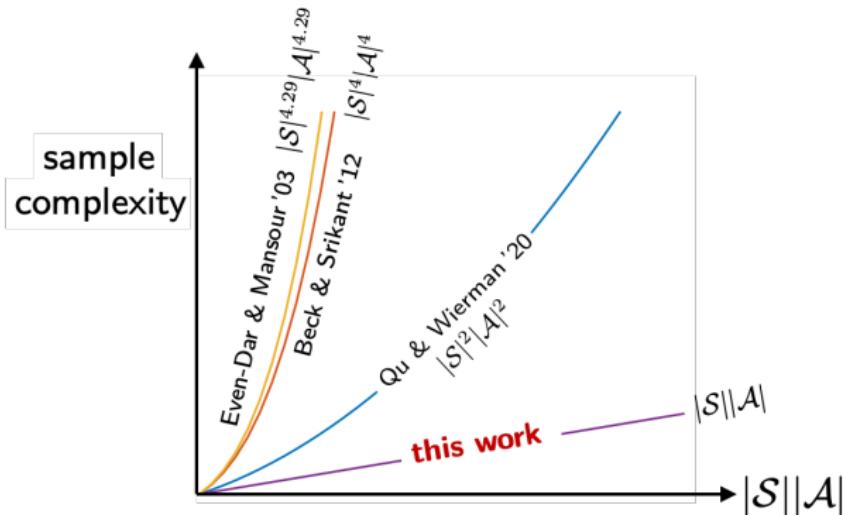
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, sample complexity of async Q-learning to yield $\|\hat{Q} - Q^*\|_\infty \leq \varepsilon$ is at most (up to some log factor)

$$\frac{1}{\mu_{\min}(1-\gamma)^5 \varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)}$$

- The first term can be improved further to $\frac{1}{\mu_{\min}(1-\gamma)^4 \varepsilon^2}$ [Li et al., 2021] for $0 < \varepsilon \leq 1$.

A collection of prior art: async Q-learning

Question: how many samples are needed to ensure $\|\hat{Q} - Q^*\|_\infty \leq \varepsilon$?

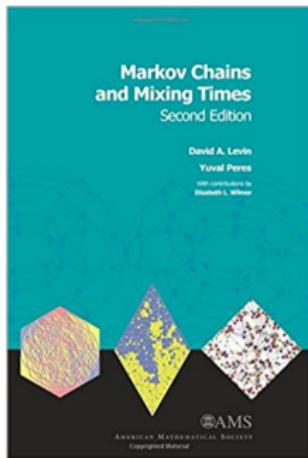


$$\text{if we take } \mu_{\min} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}, t_{\text{cover}} \asymp \frac{t_{\text{mix}}}{\mu_{\min}}$$

All prior results require sample size of at least $t_{\text{mix}}|\mathcal{S}|^2|\mathcal{A}|^2$!

Effect of mixing time on sample complexity

$$\frac{1}{\mu_{\min}(1-\gamma)^5 \varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)}$$



- reflects cost taken to reach steady state
- one-time expense (almost independent of ε)
 - it becomes amortized as algorithm runs

Minimax lower bound

minimax lower bound
(Azar et al. '13)

$$\frac{1}{\mu_{\min}(1-\gamma)^3 \varepsilon^2}$$

asyn Q-learning
(ignoring dependency on t_{mix})

$$\frac{1}{\mu_{\min}(1-\gamma)^5 \varepsilon^2}$$

Can we improve dependency on **discount complexity** $\frac{1}{1-\gamma}$?

One strategy: variance reduction

—[Wainwright, 2019b, Li et al., 2022]

Variance-reduced Q-learning updates

$$Q_t(s_t, a_t) = (1 - \eta)Q_{t-1}(s_t, a_t) + \eta \left(\mathcal{T}_t(Q_{t-1}) - \underbrace{\mathcal{T}_t(\bar{Q}) + \tilde{\mathcal{T}}(\bar{Q})}_{\text{use } \bar{Q} \text{ to help reduce variability}} \right)(s_t, a_t)$$

- \bar{Q} : some reference Q-estimate
- $\tilde{\mathcal{T}}$: empirical Bellman operator (using a batch of samples)

Variance-reduced Q-learning

—[Wainwright, 2019b, Li et al., 2022]

update variance-reduced
 \bar{Q} Q -learning



for each epoch

1. update \bar{Q} and $\tilde{T}(\bar{Q})$
2. run variance-reduced Q-learning updates

Main result: ℓ_∞ -based sample complexity

Theorem 6 ([Li et al., 2022])

For any $0 < \varepsilon \leq 1$, sample complexity for **(async) variance-reduced Q-learning** to yield $\|\hat{Q} - Q^*\|_\infty \leq \varepsilon$ is at most on the order of

$$\frac{1}{\mu_{\min}(1-\gamma)^3 \varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)}$$

- more aggressive learning rates: $\eta_t \equiv \min \left\{ \frac{(1-\gamma)^4(1-\gamma)^2}{\gamma^2}, \frac{1}{t_{\text{mix}}} \right\}$
- minimax-optimal for $0 < \varepsilon \leq 1$

References I

-  Hasselt, H. (2010).
Double Q-learning.
Advances in neural information processing systems, 23:2613–2621.
-  Kearns, M. J. and Singh, S. P. (1999).
Finite-sample convergence rates for Q-learning and indirect algorithms.
In *Advances in neural information processing systems*, pages 996–1002.
-  Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2021).
Is Q-learning minimax optimal? a tight sample complexity analysis.
arXiv preprint arXiv:2102.06548.
-  Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2022).
Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction.
IEEE Transactions on Information Theory, 68(1):448–473.
-  Thrun, S. and Schwartz, A. (1993).
Issues in using function approximation for reinforcement learning.
In *Proceedings of the Fourth Connectionist Models Summer School*, volume 255, page 263.
Hillsdale, NJ.
-  Van Hasselt, H., Guez, A., and Silver, D. (2016).
Deep reinforcement learning with double Q-learning.
In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

References II

-  Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞ -bounds for Q-learning.
arXiv preprint arXiv:1905.06265.
-  Wainwright, M. J. (2019b). Variance-reduced Q-learning is minimax optimal.
arXiv preprint arXiv:1906.04697.
-  Watkins, C. J. and Dayan, P. (1992). Q-learning.
Machine learning, 8(3-4):279–292.